doi:10.7289/V5/NCEP-ON-483

U. S. Department of Commerce
National Oceanic and Atmospheric Administration
National Weather Service
National Centers for Environmental Prediction

Office Note 483

Neural Network Technique for Gap-Filling Satellite Ocean Color Observations

Sudhir Nadiga[1], Vladimir Krasnopolsky[2], Eric Bayler[3], Hae-Cheol Kim[1], Avichal Mehra[2], and David Behringer[2]

[1]IMSG at NWS/NCEP/EMC, [2]NWS/NCEP/EMC, [3]NESDIS/STAR

August 2016

MMAB Contribution No. 328

Email: Sudhir.Nadiga@noaa.gov; Phone: 301-683-3751

Contents

Abstract

Integrating/assimilating satellite ocean color (OC) fields (chlorophyll-a, Kd_{490}, Kd_{PAR}) in NOAA's operational ocean models requires scientifically consistent and robust techniques to address temporal and spatial gaps in data, especially gaps longer than a few days. In this work, we introduce one possible approach based on a Neural Network (NN) gap-filling technique, linking OC variability, which is primarily driven by biological processes, with the physical processes of the upper ocean. A NN method for correlating satellite OC fields with other assimilated satellite and *in situ* observations: a) instigates fewer assimilation errors (since the inputs to the NN are already being assimilated) and b) reduces reliance on sparse *in situ* OC observations. In this study, satellite-derived surface variables (sea-surface temperature (SST), sea-surface height (SSH), and sea-surface salinity (SSS) fields) and gridded ARGO salinity and temperature profiles, from 0 to 75m depth, are employed as signatures of upper-ocean dynamics. Chlorophyll-a (Chl-a) fields from NOAA's operational Visible Imaging Infrared Radiometer Suite (VIIRS) are used for NN developments, as well as NOAA SSH and SST fields and NASA Aquarius mission SSS fields. The OC data correlations with the satellite SSH/SST/SSS fields are spatially and temporally dependent. The NN technique is trained using data for two years (2012 and 2013) and tested on the remaining year (2014). Results are assessed using the root-mean-square error (RMSE) and cross-correlations between observed OC fields and NN output. To reduce the impact of noise in the data while obtaining a stable computation of the NN Jacobian for sensitivity studies, an ensemble of NN with different weights is constructed. The results for the ensemble mean are compared with those for a single NN. The long-term objectives with respect to these NN efforts are to significantly expand the utility and use of satellite ocean color fields (Chl-a, Kd_{490}, Kd_{par}) by using neural network techniques to 1) develop complete and consistent global ocean color fields using satellite observations, addressing gaps, such as those resulting from swath ground track separation, obscured areas, and satellite data loss and 2) statistically predict satellite-derived ocean color fields for numerical prediction applications. This new gap-filling capability will enable the assimilation of near-real-time (NRT) ocean color data into NOAA's operational numerical modeling to address a bio-physical feedback process that is particularly important to ocean-atmosphere coupled modeling. The assimilation of ocean color data also drives/constrains modeled physical-biogeochemical processes that underlie ecological forecasting. Efforts to incorporate biogeochemical components into NCEP operational global ocean models have already commenced. While this note details preliminary work, a subsequent optimal configuration of the NN technique will be assessed using extensive experimentation, sensitivity tests, statistical metrics, using ocean modeling for validation.

List of Acronyms

ARGO	A system for observing temperature, salinity, and currents profiles in Earth’s oceans
BGC	Bio-geochemical
CC	Correlation Coefficient
CFS	Climate Forecast System
CFSR	Coupled Forecast System Reanalysis
Chl-a	Chlorophyll-a
GODAS	Global Ocean Data Assimilation System
JPSS	Joint Polar Satellite System
MLP	Multi-layer Perceptron
NN	Neural Networks
NRT	Near-Real-Time
OC	Ocean Color
RMSE	Root Mean Square Error
RTOFS	Real-Time Ocean Forecast System
SST	Sea Surface Temperature
SSH	Height
SSS	Sea Surface Salinity
VIIRS	Visible Infrared Imaging Radiometer Suite

I. Introduction

Operational integration/assimilation of operational ocean color (OC) fields (chlorophyll, Kd490, Kd_{PAR}) into ocean models has three fundamental requirements/conditions:

1. gaps in the observations, spatial and temporal, need to be filled.
2. data assimilation must be for a predicted parameter; and
3. the data being assimilated must have a long data record to facilitate compilation of a robust statistical database spanning multiple seasons.

A new approach, presented here, aims to significantly expand the utility and use of Joint Polar Satellite System (JPSS) Visible Infrared Imaging Radiometer Suite (VIIRS) satellite OC fields (chlorophyll-a (Chl-a), Kd_{490}, Kd_{PAR}) by using neural network (NN) techniques to fill spatial and temporal gaps, thereby supporting data assimilation. This approach permits:

- developing complete and consistent long-term global OC fields from satellite and *in situ* observations, addressing gaps, such as those resulting from swath ground track separation, obscured areas, and satellite data loss and
- statistically predicting satellite-derived OC fields for numerical prediction applications, enabling the assimilation of near-real-time (NRT) OC data into numerical models.
- The neural NN technique linkages between a signature of biological processes, i.e. satellite-derived OC fields, and signatures of upper-ocean physical processes, i.e. other satellite-derived and *in situ* physical variables. The optimal configuration of this preliminary NN technique setup was assessed using extensive experimentation, sensitivity tests, and statistical metrics. The final NN simulated OC will be examined using global and local statistical metrics and **validated** with global ocean model simulations.

Satellite remote-sensing of ocean color parameters provides the only means for broadly observing the biological component of the world's oceans. Consequently, this capability must be exploited for analyzing and predicting ocean bio-physical processes and establishing a linkage to biological components of ocean ecological forecasts. The assimilation of ocean color data also drives/constrains the modeling of physical-biogeochemical processes that underlie ecological forecasting. The three principal components of this OC NN effort are: a) developing an appropriate neural network for estimating ocean color parameters (chlorophyll-a, K_{490}, K_{PAR}); b) applying the ocean color neural network to create consistent ocean color fields bridging multiple satellite ocean color missions; and c) integrating a blended composite of near-real-time ocean color data and the neural network estimates (capturing low-frequency large-scale variability) into NOAA's operational ocean models. This discussion focuses on the first element, developing an appropriate gap-filling NN for estimating ocean color parameters by correlating ocean color fields to physical signatures of upper ocean processes.

I.1 Using Blended Composited Gap-Filled Ocean Color Data

I.1.1 Use in Data in Ocean and Coupled Modeling at NOAA/NCEP

The absorption of solar radiation by photosynthesis, estimated by satellite measurements of chlorophyll-a/K_{490}/K_{PAR}, can significantly modify the near-surface thermal profile. Such perturbations of the near-surface thermal profile can affect air-sea heat flux, thereby driving a bio-physical feedback to the atmosphere through changes in sea-surface temperature (SST), and is a significant input for coupled ocean-atmosphere numerical modeling predictions, particularly for longer time scales. Numerous modeling studies attest to the importance of including a solar radiation penetration-absorption scheme to better represent near-surface conditions, processes, large-scale oceanic heat transport, and coupling with the atmosphere, *e.g.* Anderson *et al.*

(2009), Ballabrera-Poy *et al*. (2007), Morel and Antoine (1994), Murtugudde *et al.* (2002), and Zhang *et al.* (2009). Ongoing efforts aim to use near-real-time VIIRS ocean color fields in NOAA's operational Real-Time Ocean Forecast System (RTOFS) (Mehra *et al*., 2011 and operational seasonal-interannual Global Ocean Data Assimilation System (GODAS; Behringer, 2007), (Saha, *et al*., 2013), the ocean component of NOAA's coupled Climate Forecast System (CFS). The work to date has established two key points: 1) the ocean responds vigorously to ocean color variability at all time scales and 2) daily VIIRS ocean color fields show significant sub-seasonal to seasonal variability. There are indications that ocean and coupled simulations are sensitive to changes in the prescribed ocean color fields, thus requiring the ocean color data stream to be of long duration, consistent, gap-filled, and free of sharp and/or persistent erroneous trends.

I.1.2 Use in biogeochemical modeling at NOAA/NCEP

The NOAA Ecological Forecasting Roadmap for 2015-2019 includes "…providing dependable, higher-quality forecast products…." Supporting this mission, a prototype foundational biogeochemical (BGC) modeling capability for NOAA's global RTOFS is being developed, aiming to use satellite ocean color data to reliably provide critical data fields for ecological forecasts.

Scientific objectives include: 1) improving NWS forecasting skill at short-term and seasonal time scales by employing coupled BGC-physical modeling framework (e.g., biological heating on the upper-ocean thermal structure); 2) investigating effects of the direct assimilation of VIIRS products (Chl-*a*) in conjunction with radiative transfer computations using an existing validated algorithm (Lee et al. 2006); 3) providing scenario-based forecasting to predict system responses to potential changes by drivers (natural or through ecosystem management decisions) using biologically and physically coupled modeling; and 4) assessing the effects of carbon flux between the atmosphere and the ocean and subsequent changes in the ocean acidity through marine biogeochemical processes.

Figure 1 depicts the three components of the coupled system: the atmospheric model, the Global Forecast System; the Global Data Assimilation System; coupled with the oceanic component, RTOFS-Global embedding biogeochemical modeling component (see the box in red color). The VIIRS ocean color observations will update and constrain both the physical (bio-physical feedback) and biogeochemical (primary productivity) components through data assimilation. The radiative transfer algorithm is a 2-band scheme, a visible band and a red band. The required inputs for this scheme are the total absorption coefficient at the surface for the 490 nm waveband (a490, m^{-1}) and the total backscattering coefficient at the surface at the same waveband (b_b490, m^{-1}). Both of these inherent optical properties are functions of chlorophyll and pure water.

The ocean physics component, RTOFS-Global runs daily, providing NCEP's "ocean weather" forecasts. The computational core of this ocean modeling system is an eddy-resolving 1/12th-degree HYbrid Coordinate Ocean Model with cylindrical and recti-linear coordinates in a Mercator projection and an Arctic bipolar patch. The latest code has 41 vertical layers which follow isopycnals in the deep ocean, z-levels near the surface, and terrain-following σ-coordinates near coastal areas. The vertical mixing scheme employs the K-Profile Parameterization (KPP; Large *et al.*, 1994). RTOFS-global is re-initialized every day using ocean initial conditions from the Navy's Coupled Ocean Data Assimilation system (Cummings, 2005).

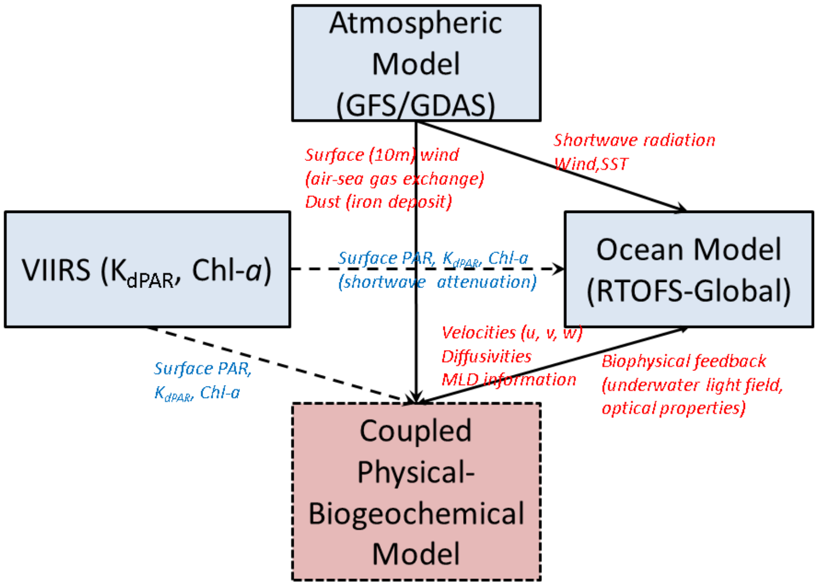


Figure 1. Schematic diagram for VIIRS ocean color data assimilation into NOAA's coupled ocean, atmosphere, data assimilation, and biogeochemical modules.

As a biogeochemical component of the coupled system, an existing well validated global BGC model (Gregg, 2002) is being coupled to RTOFS-Global. There are two sub-module components in the BGC model: ecosystem; and carbon component. Ecosystem component includes four phytoplankton groups, four nutrient groups, a single herbivore group, and three detrital pools. The phytoplankton groups are different functional types with maximum growth, sinking rates, nutrient requirements, and optical properties. Nitrate, regenerated ammonium, silica and iron are included as nutrient sources, and detritus were parameterized to simulate sinking, and remineralization of organic materials. Carbon component involves carbon cycling for dissolved organic carbon and dissolved inorganic carbon. While this biogeochemical model was originally developed for validating SeaWiFS data, our BGC model will hire data assimilative capabilities of VIIRS ocean color products in order to estimate primary productivity (photosynthetic capacity) of global ocean which indicates dynamic states of CO_2 in water and at the air-sea interface. The assimilation of VIIRS OC fields in ocean biogeochemical models will be accomplished by nudging Chl-a values in the top layer of the model towards a Ch-al analysis field. This common procedure is equivalent to the addition of a corrective Chl-a flux by nudging modeled values toward observed values. Mechanisms for Chl-a flux within water column include turbulent mixing, advection, and local biogeochemical processes, such as uptake and grazing.

I.2 NN Ocean Color Near-Real-Time Estimates

Neural network NRT ocean color estimates are aimed at: 1) gap-filling, as needed, 2) creating blended analyses in conjunction with VIIRS observations, and 3) creating an extended consistent ocean color data record spanning multiple satellite missions. While the results shown are preliminary, the NN output from the final configuration will be assessed using global and local statistical metrics and validated with observations and global ocean model simulations. So far, preliminary results indicate that the NN technique has great promise for spatially and temporally filling gaps in OC fields; therefore, we will extend our current work into statistical prediction of OC fields in the near future. The vision is a weighted blend of NN estimates and NRT VIIRS data for ocean model initialization and assimilation for i) nowcasts and one to two week ocean forecasts by NOAA’s operational RTOFS, and ii) reanalysis, establishing the best ocean initial conditions by NOAA’s operational seasonal-interannual ~~C~~coupled Climate Forecast System Reanalysis (CFSR)/GODAS. These VIIRS-based NN ocean color analyses and predictions will be assimilated when computing NRT and extended (three- to four-week) ocean and coupled weather forecasts. The concept for blending is to retain high-frequency small-scale information from the VIIRS observations while ensuring the inclusion of the low-frequency large-scale information from NN estimates. This blending methodology allows the creation of a consistent time series that spans multiple satellite missions. The NRT VIIRS data stream serves two purposes: 1) creation of a blended analysis for use by the ocean and biogeochemical models and 2) update training for the NN on a periodic basis.

Successful neural network applications for satellite remote sensing, meteorology, and oceanography have steadily increased over the last two decades (Camps-Valls and Bruzzone, 2009; Krasnopolsky, 2013). These NN applications address issues such as classification, feature extraction and tracking, pattern recognition, change detection, solving forward and inverse problems, as well as for filling missing data gaps in the measurement time series (Hidalgo *et al*., 1995; Arena and Puca, 2004; Makarynskyy and Makarynska, 2007). Peres *et al*. (2015) used NNs to extend records of observations. Multiple NN applications aim to solve forward and inverse problem in satellite OC remote sensing (Dzwonkowskia and Yan, 2005, and references there). NNs have also been applied to merge data collected by different instruments (Kwiatkowska and Fargion 2002). This work presents a new NN approach that serves to fill spatial and temporal gaps in satellite-derived OC fields by using satellite and *in situ* observations of related, but independently derived, information about the physical state of the ocean’s upper layer. Section II of this paper presents the new NN methodology. Section III describes the data used for this study. Section IV introduces the results, with,Sections V and VI providing discussion and conclusions.

II. Methodology

II.1 Neural Network background

NNs are very generic, accurate, and convenient mathematical models that emulate complicated nonlinear input/output relationships through statistical learning algorithms (Hsieh, 2009; Krasnopolsky, 2013). NNs approximate the transfer functions (mappings) between a large number of possibly-interconnected inputs and multiple outputs, even for nonlinear and not-well-known relationships. NN can be applied to any problem that can be formulated as a mapping (input vector, X, versus output vector, Y, dependence).

Mapping, *Y* (output) as a function of *X* (inputs) can be symbolically written as:

$$Y = M(X); \quad X \in \Re^n, Y \in \Re^m \tag{1}$$

where *M* denotes the mapping, *n* is the dimensionality of the input space (number of emulating NN inputs), and *m* is the dimensionality of the output space (number of emulating NN outputs). Multi-layer perceptrons (MLP) are a generic tool for approximating such mappings (Krasnopolsky, 2013). The simplest MLP NN analytical approximations use a family of functions like:

$$y_q = NN(X, a, b) = a_{q0} + \sum_{j=1}^{k} a_{qj} \cdot t_j; \;\; q = 1,2,...,m \tag{2}$$

where

$$t_j = \phi(b_{j0} + \sum_{i=1}^{n} b_{ji} \cdot x_i)$$

Here, t_j is a “neuron”, x_i and y_q are components of the input and output vectors *X* and *Y*, respectively, *a* and *b* are fitting parameters (NN weights). In the majority of NN applications the hyperbolic tangent is used as activation function ϕ. In which case, Equation (2) becomes,

$$y_q = a_{q0} + \sum_{j=1}^{k} a_{qj} \cdot \tanh(b_{j0} + \sum_{i=1}^{n} b_{ji} \cdot x_i); \quad q = 1,2,\ldots,m \tag{3}$$

In this study, Equation (3) is used for development. Equation (2 or 3) is also a mapping, which is represented symbolically as *Y = NN(X).* shown (Hsieh 2009 and Krasnopolsky 2013) that the simplest MLP (family of functions), Equation (3), can approximate any continuous, or almost continuous (with final discontinuities), mapping.

Neural networks employ adaptive weights (*a* and *b*), tuned through training with past data sets, to provide robustness with respect to random noise and fault-tolerance; thus, data sets are required to train, test, and validate the NN (Eq. 3). To train the NN, an error function (*E*) is created and minimized to find an optimal set of coefficients (a_{ij} and b_{ij}).

$$E = \frac{1}{N}\sum_{i=1}^{N}\left[Y_i - NN\left(X_i\right)\right]^2 \tag{4}$$

where N is the number of records in the training set. While neural network training is a complicated and time-consuming nonlinear optimization task, the NN training needs to be done only once for a particular application. Then, the trained NN can be repeatedly applied to new input data providing accurate and fast emulations. To retain the required accuracy, however, retraining may be required periodically. Neural networks are well-suited for parallel and vector processing.

In many practical applications (and in the application considered in this work) the mapping (1) contains an internal source of stochasticity, which may be due to: (i) errors in the data used to define the mapping, (ii) incomplete information presented to the NN by the input vector *X* (insufficient dimensionality of this vector, e.g., when some important physical parameters necessary for defining the mapping are not included), (iii) using training data with resolution too low to completely resolve the physical processes that define the mapping, etc. Therefore, in this case, the symbolic representation (1) can be modified to specifically address such uncertainty,

$$Y = M(X, \varepsilon) \tag{5}$$

where ε is a vector stochastic variable explicitly reflecting the stochastic nature of the mapping, i.e. the mapping uncertainty. Assuming that the stochastic part of the mapping is additive, representation (5) can be simplified to

$$Y = M(X) + \varepsilon \tag{6}$$

It is noteworthy that the uncertainty ε is an inherently informative part of the stochastic mapping, containing important statistical information. **Actually, the stochastic mapping is a family of mappings having a distribution function. The range and shape of the distribution function are determined by the uncertainty vector** ε. Therefore, for stochastic mapping (6), the training criterion should be modified as

$$E = \frac{1}{N}\sum_{i=1}^{N}[Y_i - NN(X_i)]^2 \le \varepsilon^2 \tag{7}$$

Thus, any NN that approximates the stochastic mapping (6) with the accuracy better than the uncertainty, ε, approximates one of the members of the family of the functions, comprising the stochastic mapping (6). Accordingly, a single NN does not adequately approximate a stochastic mapping; however, an ensemble of NNs, with all NN ensemble members satisfying the condition (7), provides an adequate approximation for the stochastic mapping (6). Several different approaches can be used to generate an ensemble of NNs that emulates the stochastic mapping (6). In this work, the ensemble is generated by giving each NN ensemble member slightly different initialization of NN weights, a_{ij} and b_{ij}, with which to start NN training. The NN ensemble members thereby correspond to different local minima of the error function (4), with all members satisfying the condition (7).

The NN Jacobian is used for evaluating sensitivities, i.e., for initial evaluation of the contributions (relative importance) of the various input variables (vector X) to the output vector Y. The NN Jacobian, J, is an $m\times n$ matrix of the first derivatives of the NN outputs over the inputs,

$$J = \left[\frac{\partial y_q}{\partial x_i}\right]_{i=1,..,n}^{q=1,...,m} \tag{8}$$

Formally speaking, the Jacobian of MLP NN (2, 3) easily can be calculated using direct differentiation,

$$\frac{\partial y_p}{\partial x_s} = \sum_{j=1}^{k} b_{pj} \cdot (1 - t_j^2) \cdot a_{js} \tag{9}$$

However, a calculation of derivatives of any statistical model (including NN) is an ill-posed problem (Krasnopolsky 2013) which should be regularized. As shown by Krasnopolsky (2007), the problem can be solved using an NN ensemble and calculating the Jacobian as an average of the Jacobians of the NN ensemble members; the approach used here.

II.2 Formulation of the problem

Ocean color parameters (chlorophyll-a/K490/KPAR) are biological, with magnitudes resulting from chains of biological processes that occur at different spatial and temporal scales within the ocean’s upper layers. The physical parameters characterizing the state of the ocean surface and upper mixed layer (SST, SSS, SSH, near-surface salinity, and near-surface temperature) constitute the active physical background for these biological processes. Thus, the variability of the physical background's spatial and temporal scales is responsible for a significant part of variability of the biological parameters. Accordingly, we consider the OC, *Y*, (the single parameter chlorophyll-a or a vector of parameters chlorophyll-a/Kd_{490}/Kd_{PAR}) to be a function, or mapping, of a vector of the ocean surface and upper-mixed-layer state variables, vector *X*.

This function is expected to be a complex nonlinear function because the variability of physical parameters is transferred into the variability of the OC through a complex hierarchy of physical, chemical, and biological processes. Also, in reality, both the OC data and the ocean state variables have finite spatial resolution (due to gridding). The physical and biological variabilities on scales finer than this resolution manifest as stochastic contributions to the OC, *Y*. Thus, the mapping between the OC, *Y*, and physical ocean variables, *X*, is a complex, nonlinear stochastic mapping (6). A stochastic variable ε represents the uncertainty introduced into the OC, *Y*, from unaccounted small (subgrid) scale variability of physical, chemical, and biological processes. The variables, constituting vectors *Y* and *X*, contain observations, which have differing levels of noise that also contribute to the stochastic variable ε.

As previously shown, a single NN does not provide an adequate emulation/approximation of a stochastic mapping (6). To effectively account for subgrid scale effects and to reduce the impact of noise in NN simulated data (e.g., the Chl-a concentration), an ensemble of NNs has been trained, using criterion (7), with the average of this ensemble serving as the simulated OC, *Y*. The Jacobian of each NN ensemble member was calculated using eq. (9) and then averaged to calculate the mean Jacobian (used in the sensitivity study below), improving the stability of the NN Jacobian.

II.3 Selecting NN inputs and outputs

Selecting an emulating NN architecture includes selecting *n* NN inputs, *m* NN outputs, and the number of hidden neurons, *k*. For this study, just one output – chlorophyll-a concentration was selected. The vector of inputs, *X*, was composed of two parts $X = \{\vec{a}, \vec{b}\}$, where $\vec{b}$ is a vector of physical parameters – sea-surface height (*SSH*), sea-surface salinity (*SSS*), sea-surface temperature (*SST*) and ARGO profiles of salinity (*sal*) and temperature (*temp*):

$$\vec{b} = \{SSH, SSS, SST, sal, temp\} \tag{10}$$

and $\vec{a}$ is a vector of auxiliary or meta variables or tracers:

$$\vec{a} = \{yr, \sin(\tau), \cos(\tau), \sin(lon), \cos(lon), \sin(lat);\ here\ \tau = 2\pi \cdot \frac{t}{366}, \tag{11}$$

where *yr* is the year, *t* is a day of the year, and *lon* and *lat* are longitude and latitude respectively. These data are included in the input vector, *X,* because for the NN is trained and applied globally over interannual periods.

For this study, input vector, *X*, defined by (10) and (11), comprises, for $\vec{a}$, all variables identified in (11) and, for $\vec{b}$, three surface parameters addressed with satellite measurements and seven near-surface layers of *sal* and *temp* from the top-75 meters of ARGO profiles. These input vector X parameters and their units, as well as the single output parameter, *Y*, are described in Table 1. All of the NNs for emulating OC mapping have 23 inputs and 1 output. Each NN (single instantiation or ensemble member) is trained to produce a simulated value of OC (Chl-a) for each grid point (latitude, longitude) based on the information available for that grid point. A global field of OC (Chl-a) is produced by sequencing through the grid points.

Table 1. A Table of Emulating Neural Network Inputs and Outputs

Input #	Variable	Units	Input	Size
1	Year		yr	1
2	Day of the year		$\sin(\frac{2 \cdot day \cdot \pi}{366 .})$	1
3	Day of the year		$\cos(\frac{2 \cdot day \cdot \pi}{366 .})$	1
4	Longitude		$\sin(lon)$	1
5	Longitude		$\cos(lon)$	1
6	Latitude		$\sin(lat)$	1
7	Sea surface height [SSH]	m	SSH	1
8	Sea surface salinity [SSS]	PSS	SSS	1
9	Sea surface temperature [SST]	°C	SST	1
10 - 16	ARGO salinity [S(z)]	PSS	sal	7
17 - 23	ARGO temperature [T(z)]	°C	temp	7
Total	All Inputs			23
Output #	Variable	Units	Output	Size
1	Chlorophyll-a	Mg/m^3	Chl-a	1

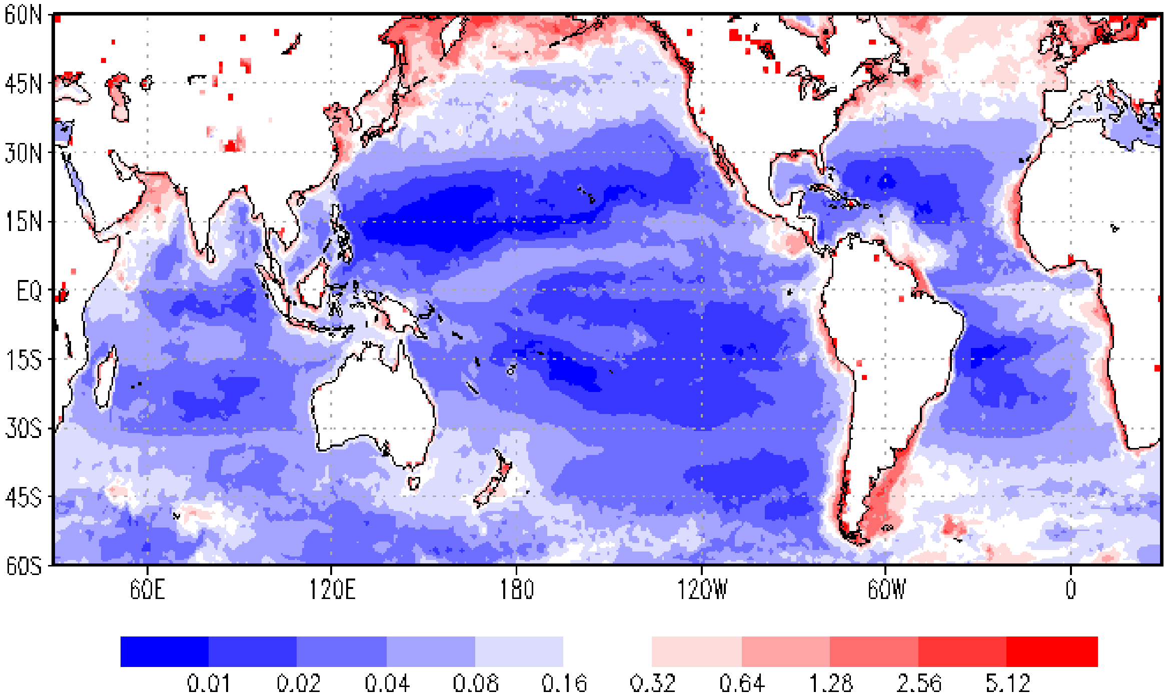
This study determines an optimal set of inputs, the parameters included in (11) and the number of near-surface layers from ARGO profiles included in (10). Approximation accuracy and the correlation coefficient between NN-generated and observed OC are used as significant indicators of each input’s importance.

III. Data

III.1 Raw data

Chlorophyll-a concentration fields from the JPSS VIIRS provide the ocean color reference data for training the neural networks in this study. Daily VIIRS chlorophyll-a global fields are compiled and interpolated from NASA's native 9-km resolution to a 1-degree latitude-longitude global grid. Figure 2 depicts the variability of the VIIRS chlorophyll-a fields.

Figure 2. Root-Mean-Square Variability of VIIRS Chlorophyll-a (mg/m^3) for 2012-2014.



The ARGO program, a collaborative international partnership program, provides upper-ocean temperature, salinity, and currents in the Earth's oceans. Global monthly-means ARGO data interpolated to daily values (Lebedev, *et al.*, 2010) from the International Pacific Research Center in Hawaii provide gridded (1-degree by 1-degree resolution) temperature and salinity profiles of the ocean's top 75m. Daily global satellite SSH data, interpolated to a 1-degree by 1-degree grid, were obtained from NOAA (Leuliette *et al.*, 2004). Daily global satellite SST, interpolated to a 1-degree by 1-degree grid, were also obtained from NOAA (Reynolds *et al.*, 2007). Satellite SSS fields, composited daily on 1-degree by 1-degree global grid were obtained from NASA's Aquarius mission (NASA JPL-PO.DAAC, Aquarius User Guide, V3, 2014; Tang *et al.*, 2014). These well-documented observations, available or soon to be available in near-real time, have been interpolated to the same global one-degree latitude-longitude grid with daily temporal resolution for the period from 2012 to 2014.

III.2 Data for NN training and validation

For NN training and validation, the first two years (2012-2013; 730 days) of daily data (approximately 20,000,000 grid point records) was selected for NN training and test. These data were split into training and test sets (approximately 10,000,000 grid point records each). Every second data record was selected for training, with those not selected being designated for testing. Each record comprises two vectors, input vector X (10, 11) and output vector Y (scalar) at each grid location and particular time (day). The 2014 data (365 days) were held for set for validating the trained NNs and estimating NN prediction (generalization) capability. For better understanding the generalization ability of the OC NN emulation, we also partitioned our data such that the 2012 data were selected for training and test sets (~ 5,000,000 grid records each), leaving both the 2013 and 2014 data for validation.

IV. Results

IV.1 The accuracy of approximation

Using the early-stopping method, this study explores the "optimal" architecture for the emulating NN, evaluates the level of uncertainty in the data, and compares performance of the single NN and NN ensemble.

IV.1.1 Selecting the number of hidden neurons

To evaluate the "optimal" size of the hidden layer, i.e., *k*, in eq. (2), a set of ten NNs, having the 23 inputs and one output identified above, were trained with *k* varying from 3 to 45. Figure 3 depicts the NN chlorophyll-a root-mean-square error (RMSE) with respect to the independent observations test set as a function of the number of hidden neurons, *k*. Figure 3 leads to the conclusion that $k = 30$ is an "optimal" number of hidden neurons, because this number provides the "best" approximation, the least RMS error, for this study's training and test sets. For $k > 30$, the NN fits to the noise in the data; thus, at least in this case, the RMSE also provides an estimate for the uncertainty, ε, in (6). Figure 3 indicates that ε is of order of 0.2 mg/m^3. This method of selecting the "optimal" size of the hidden layer is called the early-stopping method (Haykin 1994). With these considerations, $k = 30$ was selected for the NNs used in this study. Consequently, the single NNs and NN ensemble members trained and used in this study have the same architecture: 23 inputs, 30 hidden neurons, and 1 output.

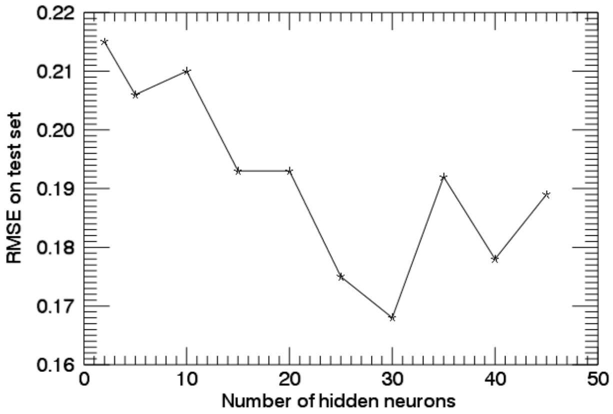


Figure 3. Neural Network chlorophyll-a root-mean-square error (RMSE; mg/m^3) versus the number of hidden neurons, *k* in Eqn (2).

IV.1.2 Estimating the value of the uncertainty

The NN RMSE has several components and can be written as,

$$RMSE = \varepsilon_{app} + \varepsilon \tag{12}$$

where ε_{app} is the approximation accuracy of NN *per se* and ε is the uncertainty due to unresolved fine-scale processes, subgrid variability, and observation errors. To better estimate the approximation accuracy obtained above ($\varepsilon < 0.2$ mg/m^3) for the NN simulating OC values, a NN was trained using the same number of hidden neurons (k = 30) and the same Chl-a output, but with only one input: the same values as the output (Chl-a), in other words, a NN emulating base configuration ((Table 2, row 2). Thus, the RMSE predominantly represents the uncertainty (approximately 0.18 mg/m^3) in the baseline 23 inputs due to subgrid processes and observation noise (Table 2, row 2). Employing a NN ensemble does not change the estimate (Table 2, row 3).

Table 2. Independent Test of Approximation Accuracy (9,209,545 records)

	Type of NN	Bias	RMSE	CC
1	**1:30:1, (ε_{app})**	4.e-5	2.e-3	1.00
2	**23:30:1**	-4.e-2	1.76e-1	0.67
3	**Ensemble 6 NNs (23:30:1)**	-2.e-2	1.72e-1	0.67
4	**24:30:1**	-4.e-2	1.73e-1	0.68
5	**23:30:1 (Chl-a ≤ 1.)**	-2.e-2	1.11e-1	0.72
6	**Ensemble (23:30:1) (Chl-a ≤ 1.)**	-2.e-2	9.1e-2	0.79
7	**24:30:1 (Chl-a ≤ 1.)**	-3.e-2	1.02e-1	0.77

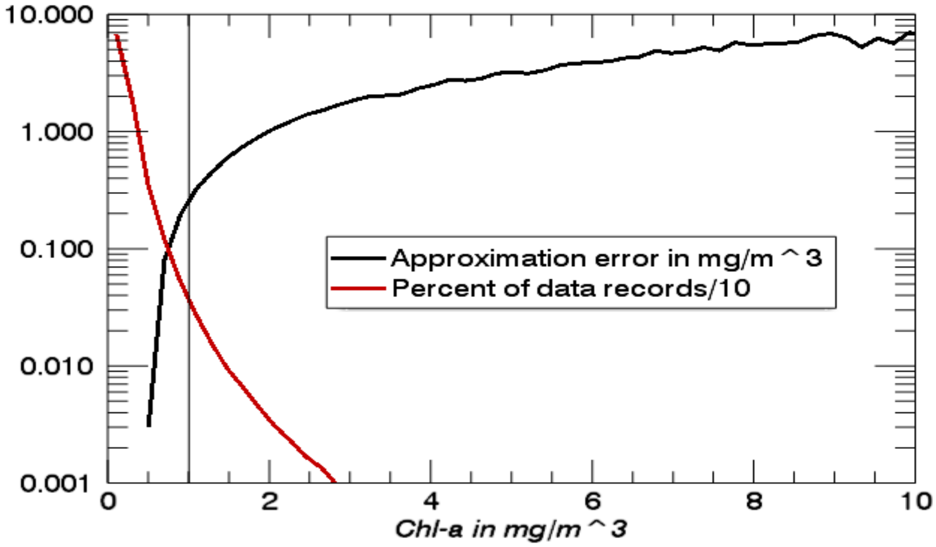


Figure 4. RMSE (black) and percentage (divided by 10) of data (red) as functions of Chl-a concentration. Vertical line shows Chl-a concentration = 1. mg/m^3.

It is clear, from physical considerations and from Figure 4, that the level of observation noise (errors) and subgrid uncertainty is higher at the higher values of Chl-a concentrations that are found mainly in coastal areas and are due to local subgrid processes. Likewise, the satellite observation errors are higher here for the accuracy of the retrieval algorithm is lower at higher levels of OC because there are only a few *in situ* observations available for the algorithm development here.

IV.1.3 Bias and RMSE

Less than one percent of the grid points in the data sets used in this study have a Chl-a concentration greater than 1.0 mg/m^3 (Fig.4); therefore, an insufficient data exists to train NNs with adequate accuracy at greater concentrations of Chl-a. Accordingly, while the NNs are trained, using the full data set; results are presented for both the full range of OC values and, when appropriate, for the case where Chl-a concentration greater than 1.0 mg/m^3 have been removed. Table 2, rows 5 through 7, show error statistics and correlation coefficients for the NN cases presented in rows 2 to 4, but with data points with Chl-a concentrations exceeding 1.0

mg/m^3 removed. When excluding the estimations for Chl-a concentration exceeding 1.0 mg/m^3, the uncertainty is significantly reduced (from order 0.17 mg/m^3 to order 0.1 mg/m^3) and the correlation between NN-simulated and observed OC is notably higher. Additionally, Table 2 highlights the significant improvement from using and NN ensemble (row 6) versus a single NN (row 5).

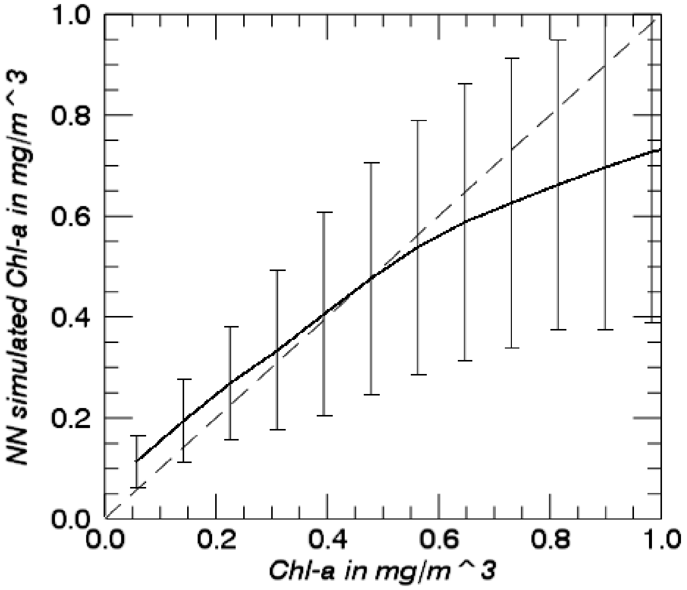


Figure 5. Binned scatter plot of NN simulated values versus observed data; bars show the standard deviation of the data within each bin.

Figure 5 shows the binned scatter plot, for Chl-a concentrations not exceeding 1.0 mg/m^3, of NN-simulated values versus observed Chl-a concentrations. Figure 6 similarly, shows the binned dependence of NN error (bias) on the magnitude of Chl-a concentration. For both figures, the bars show the standard deviation within each bin, reflecting the level of noise in the data. Figure 6 depicts a small NN negative bias for Chl-a concentrations less than 0.5 mg/m^3, with NN positive bias for Chl-a concentrations exceeding 0.5 mg/m^3; however, the magnitudes of these biases are of order of the level of the noise in the data.

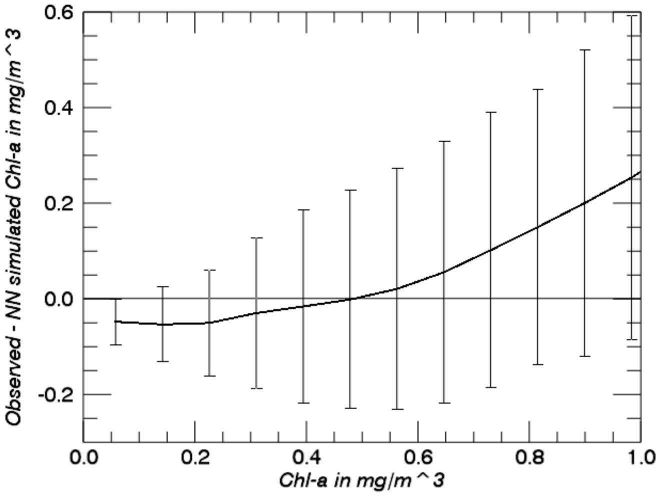


Figure 6. Binned dependence of approximation error (bias) on the value of OC; bars show the standard deviation in each bin. They reflect the level of noise in the data.

IV.1.4 Performance of NN ensemble

Finally, we trained an NN ensemble consisting of six NNs ensemble members. All six ensemble members have the same architecture 23:30:1; they were trained using different initial values for NN weights, a_{ij} and b_{ij}, in (3). Thus, different NN ensemble members correspond to different local minima of the error function (7). All NN ensemble members were trained on full NN training set. The ensemble members and the ensemble average performances are shown in Table 3 for Chl-a concentration $\leq 1.$ mg/m$^{3.}$

Table 3. Performances of NN ensemble members and NN ensemble for Chl-a concentration $\leq 1.$ mg/m^3.

Ensemble Member #	RMSE (mg/m^3)	Correlation Coefficient
1	0.11	0.722
2	0.093	0.766
3	0.097	0.757
4	0.097	0.757
5	0.094	0.758
6	0.094	0.758
Ensemble Mean	**0.091**	**0.792**

Table 3 shows that the ensemble mean has higher cross-correlation between the NN member estimated values and VIIRS observations, as well as lower RMSE than any of the individual ensemble members. The ensemble mean clearly outperforms each of the individual ensemble members, suggesting that random noise may be contaminating the input and/or the validation observation data.

IV.2 Evaluation of NN prediction

IV.2.1 Prediction accuracy

The 365 days of 2014 were used as an independent data set for validating the trained NNs. The trained NNs and the NN ensemble were applied to these data to produce Chl-a fields. Statistics were calculated comparing the NN fields to observed VIIRS Chl-a fields. Figure 7 presents NN global men RMSE time series, depicting the time series for a single NN with dashed lines and NN ensembles with solid lines. This figure supports the conclusion that the major contributors of noise in the VIIRS Chl-a data are those data with Chl-a concentrations exceeding 1.0 mg/m^3.

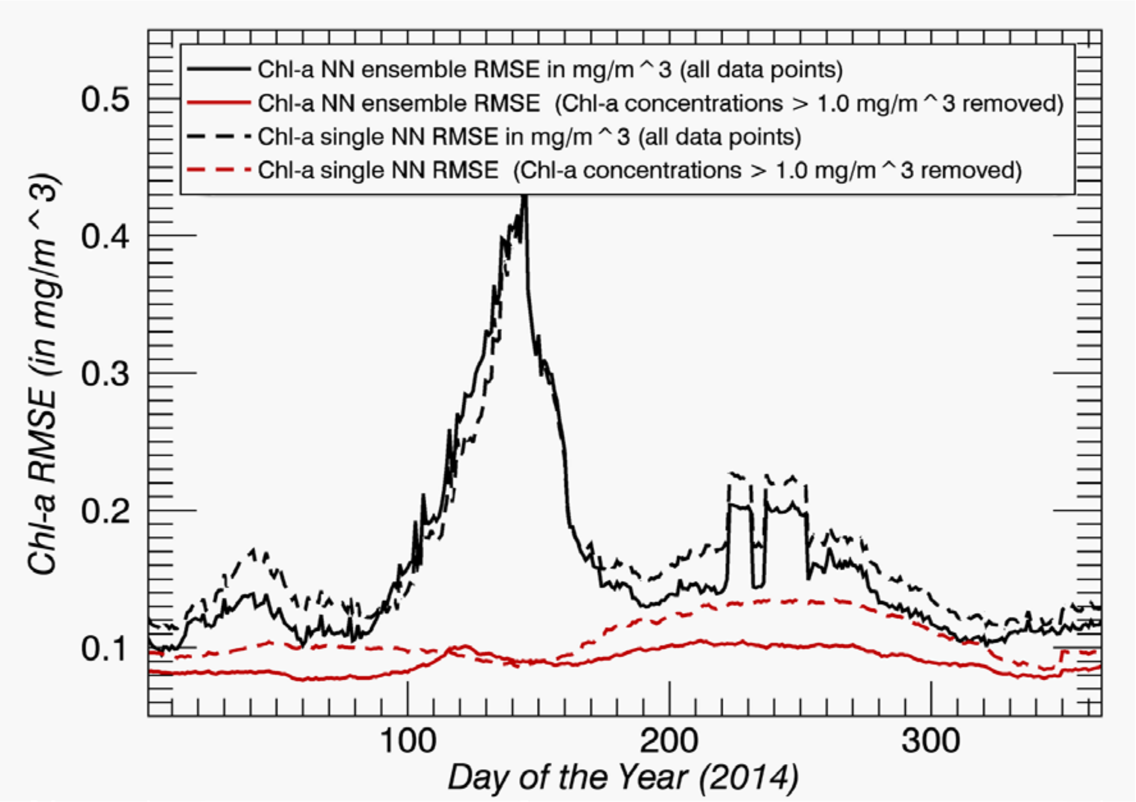


Figure 7. Neural network chlorophyll-a global mean RMSE time series: BLACK – full data set; RED – Chl-a values exceeding 1.0 mg/m^3 removed (less than 1% of data removed); SOLID lines indicate ensemble means and DASHED lines represent the mean for a single NN.

These data points, representing less than 1% of the entire data set, may create problems for retrieval algorithm development and make training NN for Chl-a concentrations greater than 1.0 mg/m^{3}difficult. The very-small mean bias seen in Figure 8 is below our estimate for the level of noise in the data (Table 2, row 7). The global mean bias for a single-NN (Figure 8) depicts a clear seasonal cycle, with positive values during the boreal winter and negative values during the austral winter.

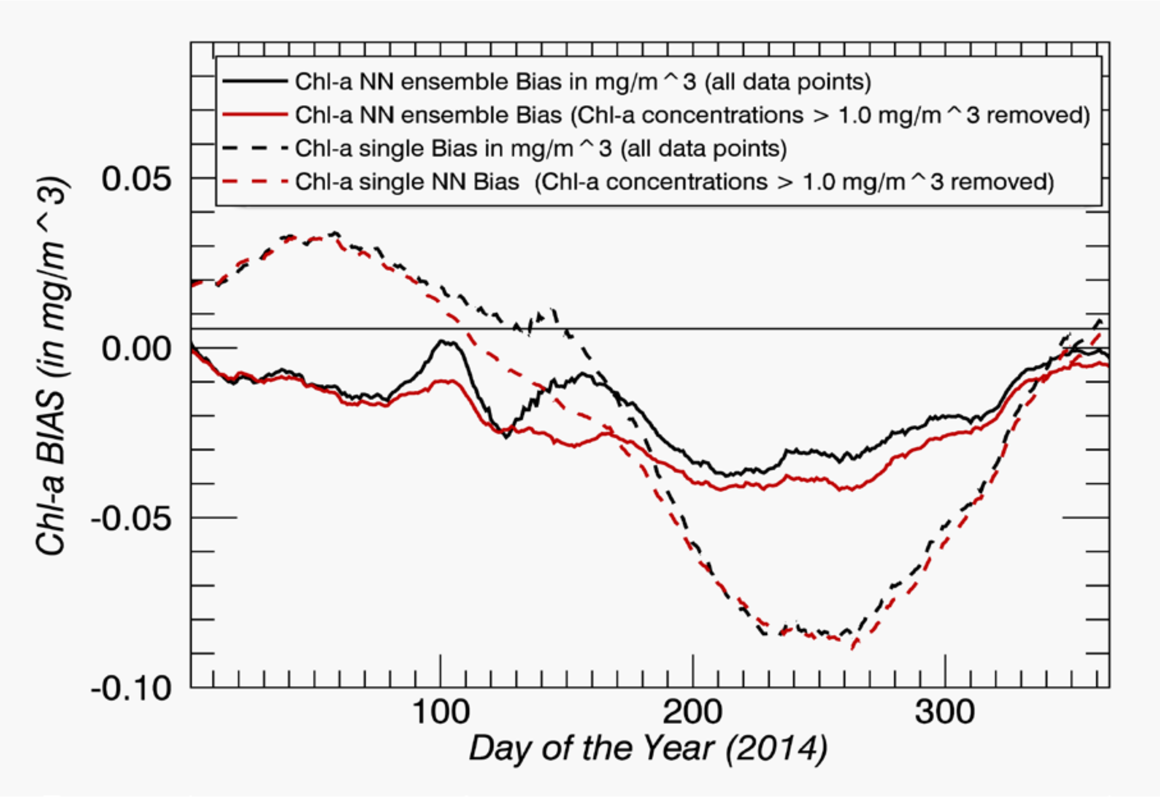


Figure 8. Neural network chlorophyll-a global mean bias time series, referenced to VIIRS observations (VIIRS – NN values): BLACK – full data set; RED – Chl-a values exceeding 1.0 mg/m^3 removed (less than 1% of data removed); SOLID lines indicate ensemble means and DASHED lines represent the mean for a single NN.

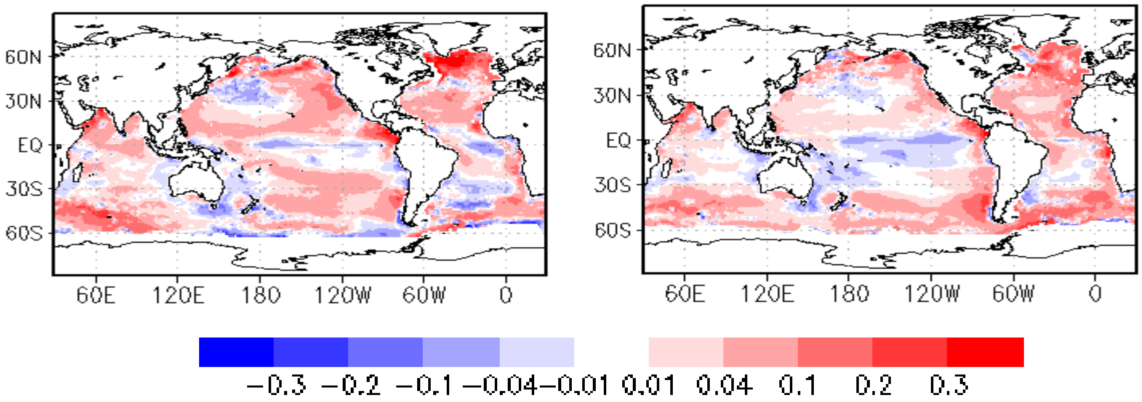


Figure 9. Global bias (Chl-a > 1 mg/m^3 removed) for a single NN (left) and for the ensemble mean (right).

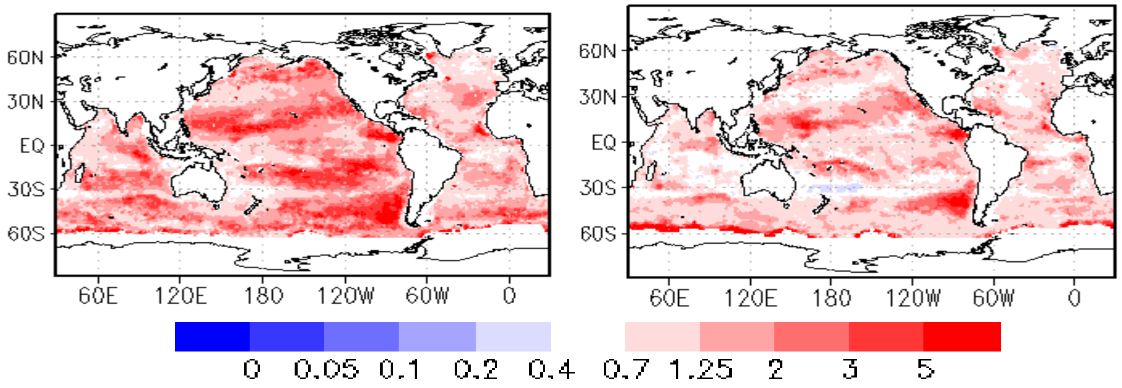


Figure 10. RMS variability of the neural network chlorophyll-a (Chl-a > 1. mg/m^3 removed) for a single NN (left) and for the ensemble mean (right).

The spatial pattern of bias (NN minus VIIRS) shown in Figure 9 has large values in high latitudes and in shallow waters, i.e., over continental shelves and in coastal regions, etc. For the ensemble mean, the overall error is small (< 0.1 mg/m^3) for most of the year; however, since the signal is small, the error to variability ratio is low only in the center of the major ocean gyres (not shown).

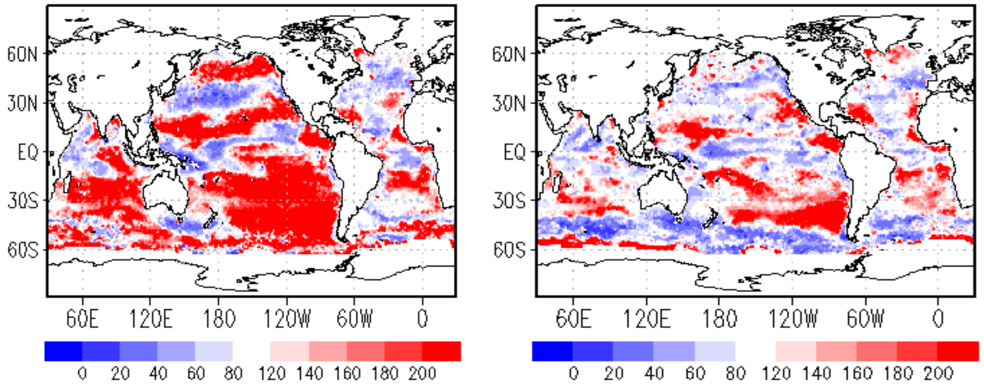


Figure 11. Variability ratio in percent: neural network chlorophyll-a variability (Chl-a > 1 mg/m^3 removed) divided by variability of VIIRS observations for a single NN (left) and for the ensemble mean (right).

Figure 11 portrays that the NN method successfully captures chlorophyll-a variability in the VIIRS observations. The single NN estimates are over-energetic with respect to VIIRS observations, while the ensemble mean has approximately the same level of variability as the VIIRS observations. Typically, the NN is over-energetic compared to the VIIRS observations in the oligotrophic subtropics, where the mean Chl-a values are very low. In these regions, the signal-to-error ratio is low, and the NN technique has difficulty in reproducing the observed variability. We hypothesize that it may be necessary to retrain a different set of NNs specifically for these regions where the observed variability of Chl-a is weak. This task is beyond the scope of this preliminary study and will be taken up later. It is clear that random error contaminates the Chl-a observations and inputs and the NN ensemble is able to remove some of these errors. Collectively, Figures 7 through 11 demonstrate that the NN ensemble significantly outperforms a single NN.

The mean cross-correlation (CC) of neural network estimates with VIIRS observations (Figure 12) is relatively stable and high throughout the validation period (~ 0.8), which is reassuring. The CC is lower and more variable for the case where all data points are retained, suggesting that the few data points with Chl-a concentrations exceeding 1 mg/m^3 are responsible for notably degrading NN performance. The NN has difficulty in certain regions, possibly due to inadequate sampling of the inputs (SST, SSS, T, and S) in regions with large spatial gradients and, especially, notable temporal variability in VIIRS-observed Chl-a values. The ensemble mean has higher CC with VIIRS observations than does a single NN (Figure 13), predominantly in the mid-latitude North Pacific and North Atlantic oceans.

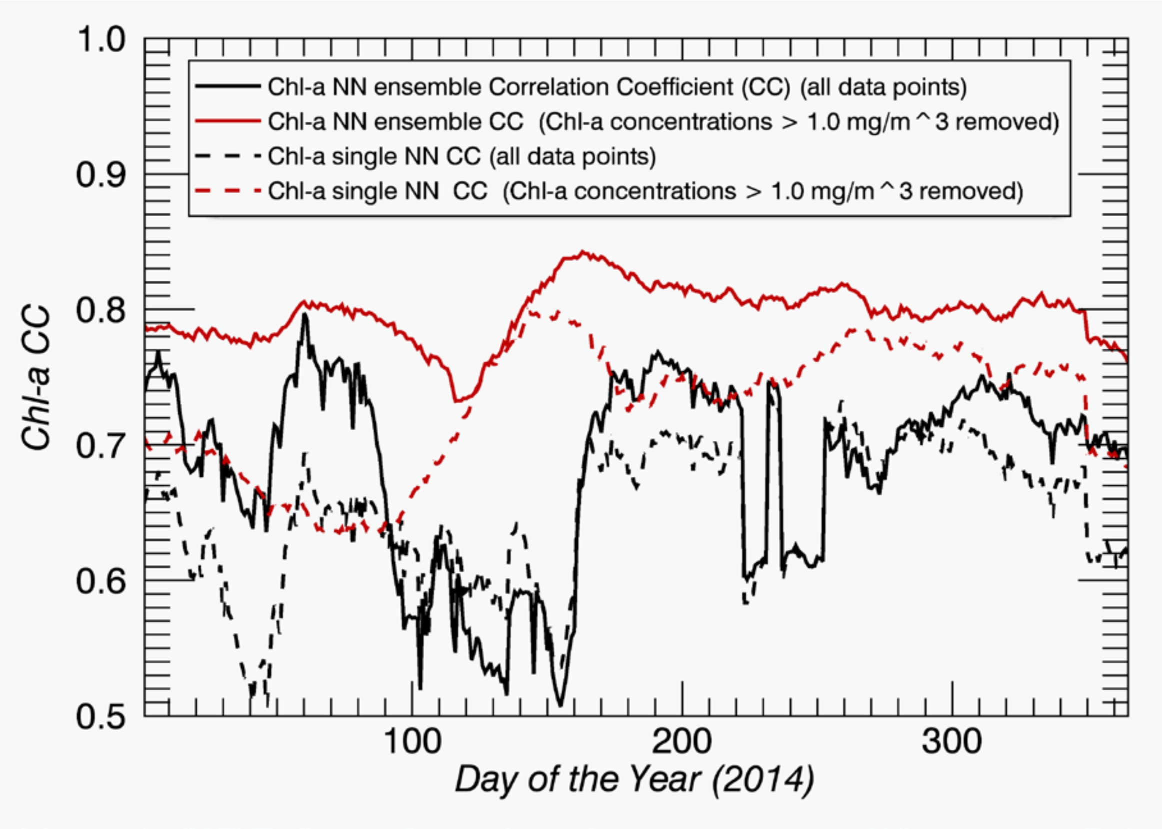


Figure 12. Neural network chlorophyll-a cross-correlation with VIIRS observations: BLACK – full data set; RED – Chl-a values exceeding 1.0 mg/m^3 removed (less than 1% of data removed); SOLID lines indicate ensemble means and DASHED lines represent the mean for a single NN.

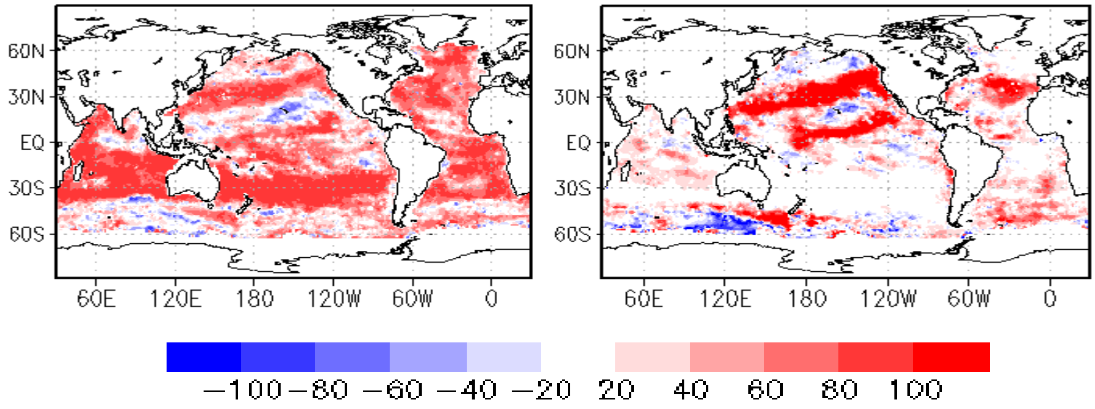


Figure 13. Spatial plot of global cross-correlation (Chl-a > 1mg/m^3 removed) in percent for ensemble mean (left)and cross-correlation difference between ensemble mean and single NN (right).

IV.2.2 Preliminary evaluation of NN sensitivity

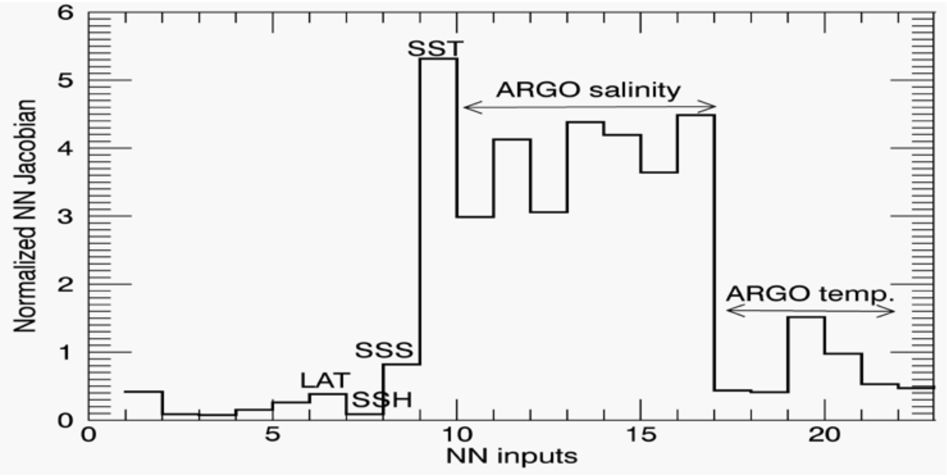


Figure 14. Linearized estimates, employing the neural network Jacobian, of the contributions from the various input parameters (latitude, daily SSH, daily SSS, daily SST, monthly ARGO surface and subsurface temperature and salinity) to the neural network output.

The NN Jacobian enables an initial evaluation of the sensitivity of the output chlorophyll-a vector, *Y*, to the contribution of the various parameters included in the input vector *X*, *i.e.*, the relative importance of the input parameters. After calculating the NN Jacobian (Eqn. 9) for each NN ensemble member; the mean Jacobian was calculated. Figure 14 plots the absolute value for each component of the mean NN Jacobian vector, providing a linear approximation of the importance of each NN input parameter. Figure 14 demonstrates that **the most important input parameter is SST, followed by the ARGO salinity fields for the surface and subsurface**. ARGO "surface" observations of temperature and satellite sea-surface salinity are less important than those at greater depth, likely because satellite SST and ARGO surface salinity already capture portions of the surface variability of temperature and salinity, respectively.

V. Discussion and Conclusions

This study introduced a new neural-network-based approach for relating a biological parameter, chlorophyll-a concentration, to physical processes of the upper ocean. This NN maps satellite-derived surface variables (SST, SSH, and SSS fields), along with ARGO *in situ* salinity and temperature near-surface profiles), to the Chl-a concentration, effectively establishing **a NN-based empirical (statistical) biological model for Chl-a**. Chlorophyll-a (an ocean color field) from NOAA's operational VIIRS, and NOAA SSH and SST fields and NASA Aquarius mission SSS fields, spatially and temporally averaged to daily gridded (1° × 1° latitude/longitude) fields, were used for the NN training. The NNs were trained using data for two years (2012-2013) and tested on the remaining year (2014). Results were assessed using the mean error (bias), RMSE, and cross-correlations between NN generated Chl-a values and observations. An ensemble of NNs with different weights was employed to reduce the impact of the noise in the data and to calculate the NN Jacobian for sensitivity studies,

Coarse spatial and temporal data resolution limits the features that can be observed in NN-generated Chl-a fields. As shown, global and mesoscale features are represented sufficiently well in the NN-estimated Chl-a fields; however, in order to generate finer-scale features, the NN needs to be trained on data with finer resolutions.

This study demonstrates that the NN technique provides an accurate, computationally cheap method for filling spatial and temporal gaps in satellite observation fields and time series. It is noteworthy that a single NN (or a single NN ensemble) is capable of generating OC fields on the entire global grid. The accuracy of NN prediction did not deteriorate during the validation period, performing well during the entire year (2014). These results demonstrate significant NN ability in terms of spatial and temporal generalization.

The tested method accurately estimates the seasonal cycle and large-scale spatial patterns in VIIRS Chl-a fields, best reproducing VIIRS Chl-a variability in the mid-latitudes of the major ocean gyres. The largest errors are found in areas where spatial scales of variability are small and the variability is large, e.g., continental shelves, coastal regions, marginal seas, etc. In these regions, VIIRS Chl-a variability is high and the Chl-a and other satellite-derived data have the highest levels of noise. Thus, removing data points where Chl-a exceeds 1.0 mg/m^3 (less than 1% of points) improves NN performance through the reduction of noise in input and output data. Disregarding the noisiness of the data in these regions, the amount of data in these regions is very small and not sufficient for NN training here.

The NN approach successfully eliminates the systematic component of the noise (bias). For reducing the random component of the noise, an NN ensemble was trained. As shown, the ensemble mean outperforms each of the ensemble members. Clearly, random noise must be contaminating the observation data streams.

The mean Jacobian of the NNs was used to evaluate the relative importance of NN inputs. The results show that daily SST is the most important input, closely followed by ARGO monthly near-surface salinity profiles. The ARGO monthly temperature near-surface signal moderately contributes to NN performance.

The vision, supported by this study, is a weighted blend of NN estimates and NRT VIIRS data for ocean model initialization and assimilation for:

- nowcasts and one-to-two week ocean forecasts by NOAA’s operational RTOFS, and
- reanalysis, establishing the best ocean initial conditions with NOAA’s operational seasonal-interannual CFSR/GODAS.

These VIIRS-based NN OC analyses and predictions will be assimilated when computing NRT and extended (three- to four-week) ocean and coupled weather forecasts. The concept is to retain high-frequency small-scale information using the VIIRS observations while ensuring the inclusion of the low-frequency large-scale information in the NN estimates. This blending methodology will allows the creation of a consistent time series that spans multiple satellite missions. The NRT VIIRS data stream will serves two purposes: 1) creation of a blended analysis for use by the ocean and biogeochemical models and 2) update training for the NN on a periodic basis. NN skill will improve by: 1) optimizing NN inputs, 2) retraining NN with accumulated new data, 3) introducing additional information (additional NN inputs and outputs), and 4) finer resolutions (better than 1°x 1° latitude/longitude) for both inputs and outputs.

Our future plans include investigation of the following topics:

1. Is the particular set of chosen physical input parameters minimal for all areas of the World Oceans?
2. Is there a reason to include incoming SWR as an input parameter of interest?
3. Does it make sense to derive simultaneously of all three OC parameters (Chl-a, Kpar, and K950).

References:

Anderson, W., Gnanadesikan, A., and Wittenberg, A., 2009. Regional impacts of ocean color on tropical Pacific variability, *Ocean Sci., 5,* 313-327.

Arena, F, Puca S, 2004. The reconstruction of significant wave height time series by using a neural network approach. *Journal of Offshore Mechanics and Arctic Engineering*, 126 (3), 213-219

Ballabrera-Poy, J., R. Murtugudde, R.-H Zhang, and A. Busalacchi, 2007. Coupled Ocean-Atmosphere Response to Seasonal Modulation of Ocean Color: Impact on Interannual Climate Simulations in the Tropical Pacific, *J. Clim., 20,* 353-374.

Behringer, D.W., 2007. The global ocean data assimilation system at NCEP. 11th symposium on integrating observing and assimilation systems for atmosphere, oceans, and land surface, AMS 87th annual meeting, San Antonio, TX, 12pp.

Camps-Valls G, L Bruzzone, 2009. Kernel methods for remote sensing data analysis, - Wiley Online Library.

Cummings, J. A., 2005. Operational multivariate ocean data assimilation. *Q. J. Roy. Meteor. Soc.*, 131, 3583–3604

Dzwonkowskia, B., and X.-H. Yan, 2005. Development and application of a neural network-based ocean colour algorithm in coastal waters, *International Journal of Remote Sensing,* Volume 26, Issue 6, pages 1175-1200

Gregg, W., 2002. Tracking the SeaWiFS record with a coupled physical/biogeochemical/radiative model of the global oceans, *Deep Sea Res.* II, 49, 81-105

Haykin S., 1994. Neural Networks: A Comprehensive Foundation. Macmillan College Publishing Company, New York, USA

Hidalgo, O. S., J. C. Nieto OMAE 1995. Filling missing observations in time series of significant wave height. 14th Intl Conf on Offshore Mechanics & Arctic Engng; 18-22 June 1995; Copenhagen, Denmark. Sponsored by ASME et al. Procs. Publ by ASME, ISBN 0-7918-1308-8. Vol II, pp. 9-17

Hsieh W.W., 2009. Machine Learning Methods in the Environmental Sciences. Cambridge University Press, Cambridge

Krasnopolsky V., 2007. “Reducing Uncertainties in Neural Network Jacobians and Improving Accuracy of Neural Network Emulations with NN Ensemble Approaches”, *Neural Networks*, 20, pp. 454–461

Krasnopolsky V., 2013. "The Application of Neural Networks in the Earth System Sciences. Neural Network Emulations for Complex Multidimensional Mappings", Springer, 200 pp

Kwiatkowska, E., J.Fargion, G.S., 2002. Merger of ocean color information from multiple satellite missions under the NASA SIMBIOS Project Office. Proceedings of the Fifth International Conference on Information Fusion, 2002, Year: 2002, Volume: 1, Pages: 291 - 298 vol.1, DOI: 10.1109/ICIF.2002.1021164

Large, W.C., J.C. McWilliams and S.C. Doney, 1994. Oceanic vertical mixing: a review and a model wit a nonlocal boundary layer parameterization. *Rev. Geophys*. 32, 363-403.

Lebedev, K. V., S. DeCarlo, P. W. Hacker, N. A. Maximenko, J. T.Potemra, and Y. Shen, 2010. Argo Products at the Asia-Pacific Data-Research Center, Eos Trans. AGU, 91(26), Ocean Sci. Meet. Suppl. Abstract IT25A-01.

Lee, Z.P., D. KePing, A. Robert, L. SooChin and B. Penta, 2006. Penetration of solar radiation in the upper ocean: A numerical model for oceanic and coastal waters. *J. Geophys. Res*., 110, 1-12.

Leuliette, E. W., et al., 2004. Calibration of TOPEX/Poseidon and Jason Altimeter Data to Construct a Continuous Record of Mean Sea Level Change, *Marine Geodesy*, 27(1-2), 79-94.

Makarynskyy, O., D. Makarynska, 2007. Wave prediction and data supplementation with artificial neural network. *Journal of Coastal Research*, 23 (4), 951-960.

Mehra et al., 2011. A Real Time Operational Global Ocean Forecast System. US GODAE OceanView Workshop on Observing System Evaluation and Intercomparisons, Univ. of California Santa Cruz, CA, USA, 13-17 June 2011.
(available at http://polar.ncep.noaa.gov/global/about/GODAE11_poster_d1.pdf)

Morel, A., and D. Antoine, 1994. Heating rate within the upper ocean in relation to its bio-optical state, *J. Phys. Oceanogr., 24,* 1652-1665.

Murtugudde, R., Beauchamp, J., McClain, C.R., Lewis, M., and Busalacchi, A., 2002, Effects of penetrative radiation on the upper tropical ocean circulation, *J. Clim., 15,* 470-486.

Peres D.J., C. Iuppa, L. Cavallaro, A. Cancelliere, E. Foti, 2015. Significant wave height record extension by neural networks and reanalysis data, Oean Moddeling (in press).

Reynolds, R. W., T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax, 2007. Daily high-resolution blended analyses for sea surface temperature. *J. Climate*, 20, 5473-5496.

Saha, S., and co-authors, 2013, The NCEP climate forecast system version 2, *J. Clim.*,doi:, 10.1175/JCLI-D-12-00823.1

Tang, Wenqing, et al., 2014. “Validation of Aquarius sea surface salinity with in situ measurements from Argo floats and moored buoys”, *J. Geophys. Res. (Oceans)*, Vol. 119, Issue 9.

Zhang, R.-H.,, Busalacchi, A., Wang, X., Ballabrera-Poy, J., Murtugudde, R., Hackert, E., Chen, D., 2009. Role of ocean biology-induced feedback in the modulation of El Nino-Southern Oscillation, *Geophys. Res., Lett.*, ., 36, L03608.